Lattice QCD: New Computing Cluster (“Ds”)

Don Holmgren

CD/SCF/HPC

All Experimenters Meeting, November 1, 2010

Context

- The High Performance Computing Department of Fermilab's Computing Division designs, procures, and operates computer clusters dedicated to Lattice QCD computations as part of the DOE SC LQCD-ext project
  - $19.2M over 5 years (FY10-FY14), $14M of the total at FNAL, a continuation of the 4-year, $9.2M DOE SC LQCD project (FY06-FY09), funded by DOE NP and HEP
  - Three labs: FNAL, Jefferson Lab, Brookhaven
  - FNAL operates 60% of the capacity (will be 75% once "Ds" is included)
  - Fermilab personnel: Bill Boroski (Contractor Project Manager), Bakul Banerjee (Associate Contractor Project Manager), Jim Simone (Dept Head), Amitoj Singh (Deputy Dept Head), Don Holmgren, Nirmal Seenu, Bob Forster, Rick van Conant, Ken Schumacher, Kurt Ruthmansdorfer
- DOE SC (HEP, NP, ASCR) also funds LQCD software development through the SciDAC-2 (Scientific Discovery thru Advanced Computing) program (2006-2011)
  - First SciDAC grant (2002-2005) funded prototype clusters at FNAL, JLab

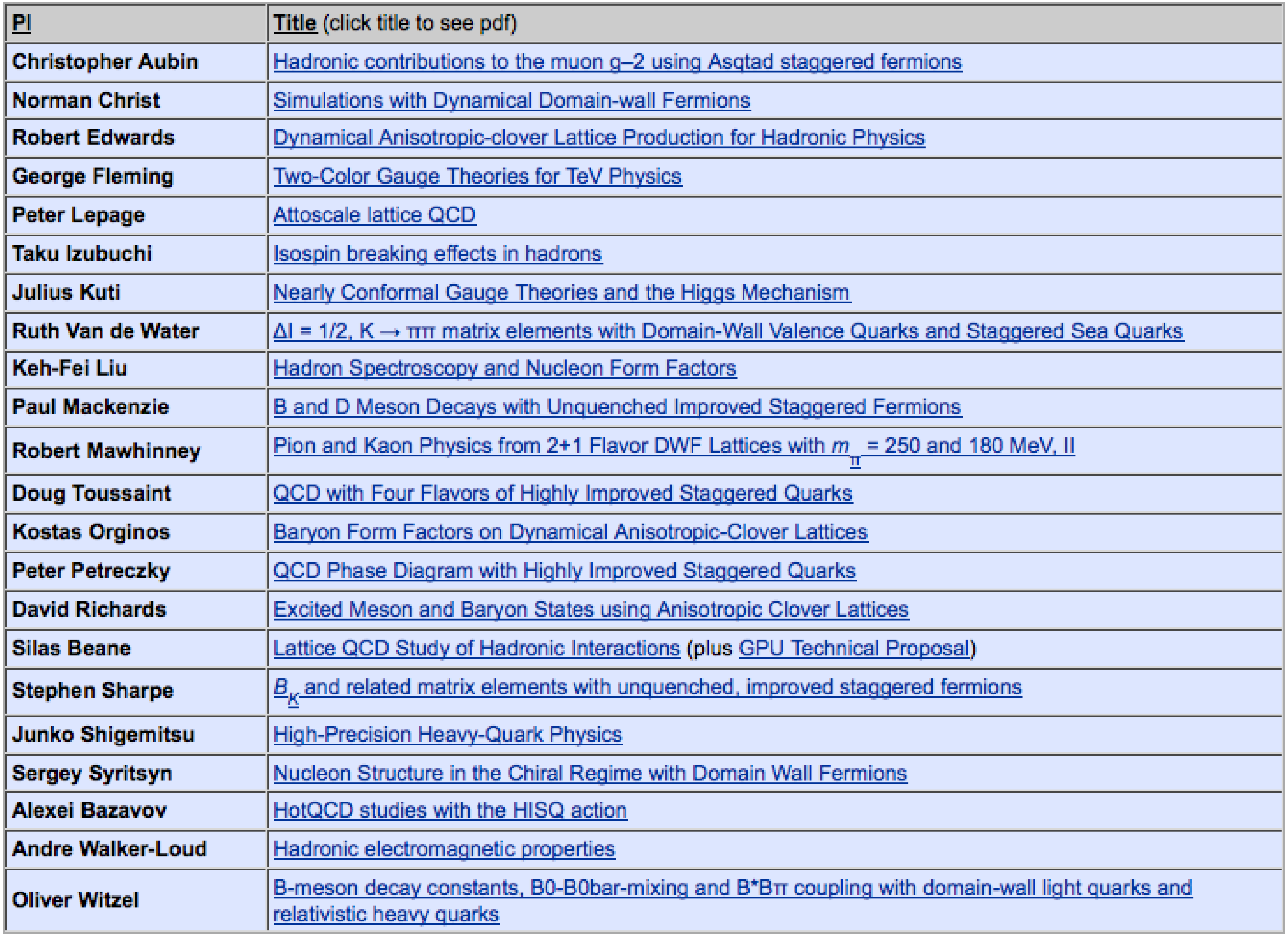
- Computing resources on LQCD-ext machines are allocated annually to USQCD (collaboration of lattice theorists) members by a national scientific program committee based on physics proposals

- Fermilab’s Paul Mackenzie is chair of the USQCD Executive Committee

- USQCD also applies for resources at DOE Leadership Computing Facilities (ANL, ORNL)
  - Part of the USQCD’s workload requires the very large “capability” machines at ANL (IBM BlueGene) and ORNL (Cray)
  - A larger part of the workload requires medium-range “capacity” machines like those at Fermilab and Jefferson Lab

- For more information, see http://www.usqcd.org/

Type A Proposals—2010



PI	Title (click title to see pdf)
Christopher Aubin	Hadronic contributions to the muon g–2 using Asqtad staggered fermions
Norman Christ	Simulations with Dynamical Domain-wall Fermions
Robert Edwards	Dynamical Anisotropic-clover Lattice Production for Hadronic Physics
George Fleming	Two-Color Gauge Theories for TeV Physics
Peter Lepage	Attoscale lattice QCD
Taku Izubuchi	Isospin breaking effects in hadrons
Julius Kuti	Nearly Conformal Gauge Theories and the Higgs Mechanism
Ruth Van de Water	$\Delta I = 1/2$, $K \to \pi\pi$ matrix elements with Domain-Wall Valence Quarks and Staggered Sea Quarks
Keh-Fei Liu	Hadron Spectroscopy and Nucleon Form Factors
Paul Mackenzie	B and D Meson Decays with Unquenched Improved Staggered Fermions
Robert Mawhinney	Pion and Kaon Physics from 2+1 Flavor DWF Lattices with m_π = 250 and 180 MeV, II
Doug Toussaint	QCD with Four Flavors of Highly Improved Staggered Quarks
Kostas Orginos	Baryon Form Factors on Dynamical Anisotropic-Clover Lattices
Peter Petreczky	QCD Phase Diagram with Highly Improved Staggered Quarks
David Richards	Excited Meson and Baryon States using Anisotropic Clover Lattices
Silas Beane	Lattice QCD Study of Hadronic Interactions (plus GPU Technical Proposal)
Stephen Sharpe	B_K and related matrix elements with unquenched, improved staggered fermions
Junko Shigemitsu	High-Precision Heavy-Quark Physics
Sergey Syritsyn	Nucleon Structure in the Chiral Regime with Domain Wall Fermions
Alexei Bazavov	HotQCD studies with the HISQ action
Andre Walker-Loud	Hadronic electromagnetic properties
Oliver Witzel	B-meson decay constants, B0-B0bar-mixing and B*Bπ coupling with domain-wall light quarks and relativistic heavy quarks

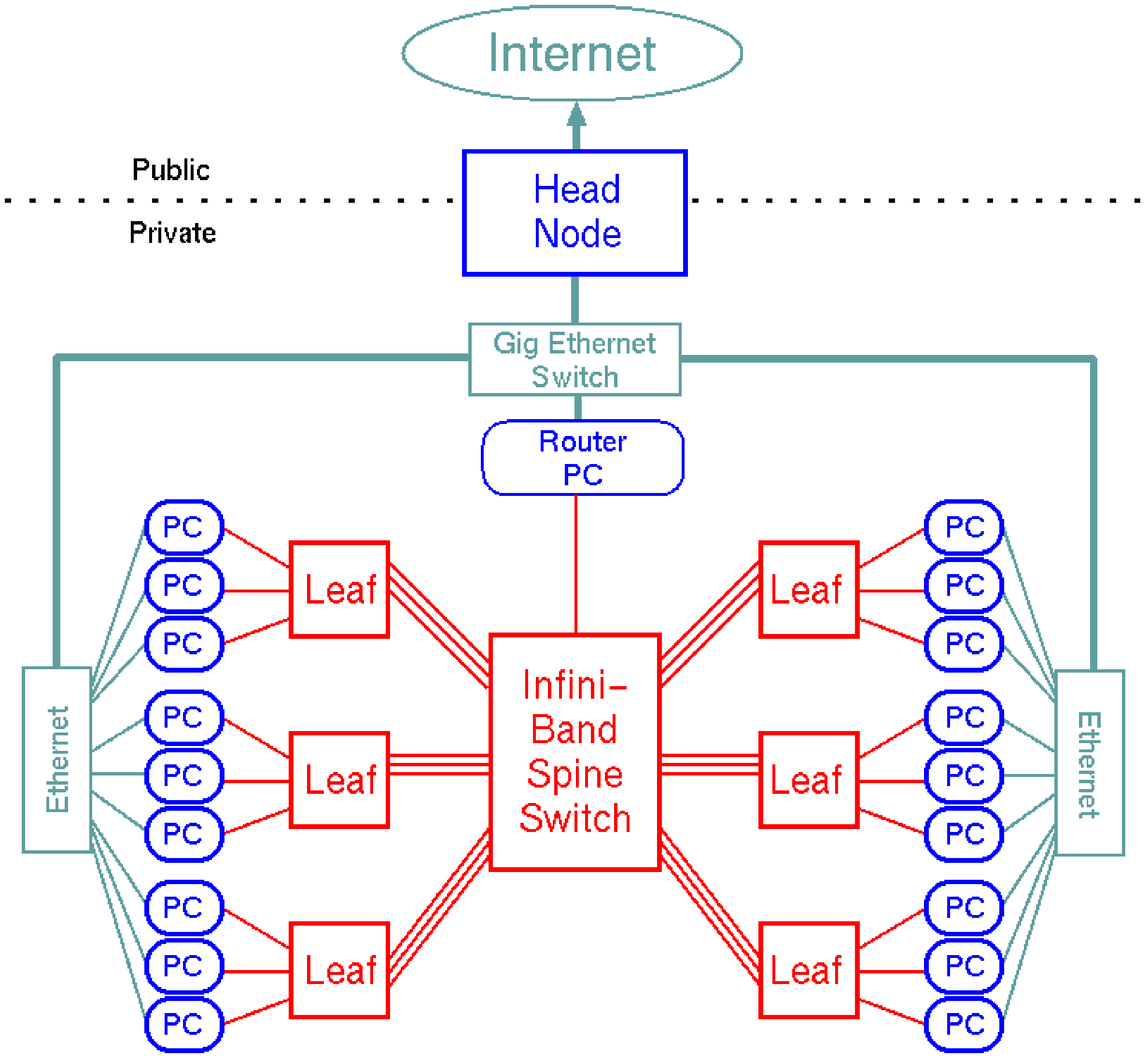
LQCD Cluster Designs

- Individual LQCD simulations require the combined power of hundreds to thousands of processors
  - Unlike reconstruction, where processors can work independently on different events, LQCD simulations require the processors to work in conjunction
  - Requirements:
    Good floating point rates (giga- to teraflops per second)
    High memory bandwidth (Gbytes/sec per processor)
    Low latency and high bandwidth communications (microsecond message latencies, 10+ Gbit/sec per processor)
  - Satisfied by:
    Commodity servers running Linux on AMD or Intel processors
    Infiniband networking hardware (10 to 40 Gbit/sec)
    Parallel programming methodologies (MPI)

Fermilab LQCD Clusters

Cluster	Nodes	Type	Cores	Peak TFlops	Sustained TFlops	Location
Kaon (FY06)	600	Dual Socket Dual Core	2400	19.2	2.6	LCC (New Muon)
J/Psi (FY08/09)	856	Dual Socket Quad Core	6848	57.5	8.4	GCC (Wideband)
Ds (FY10)	246	Quad Socket Eight Core	7872	63.0	12.5	GCC (Wideband)

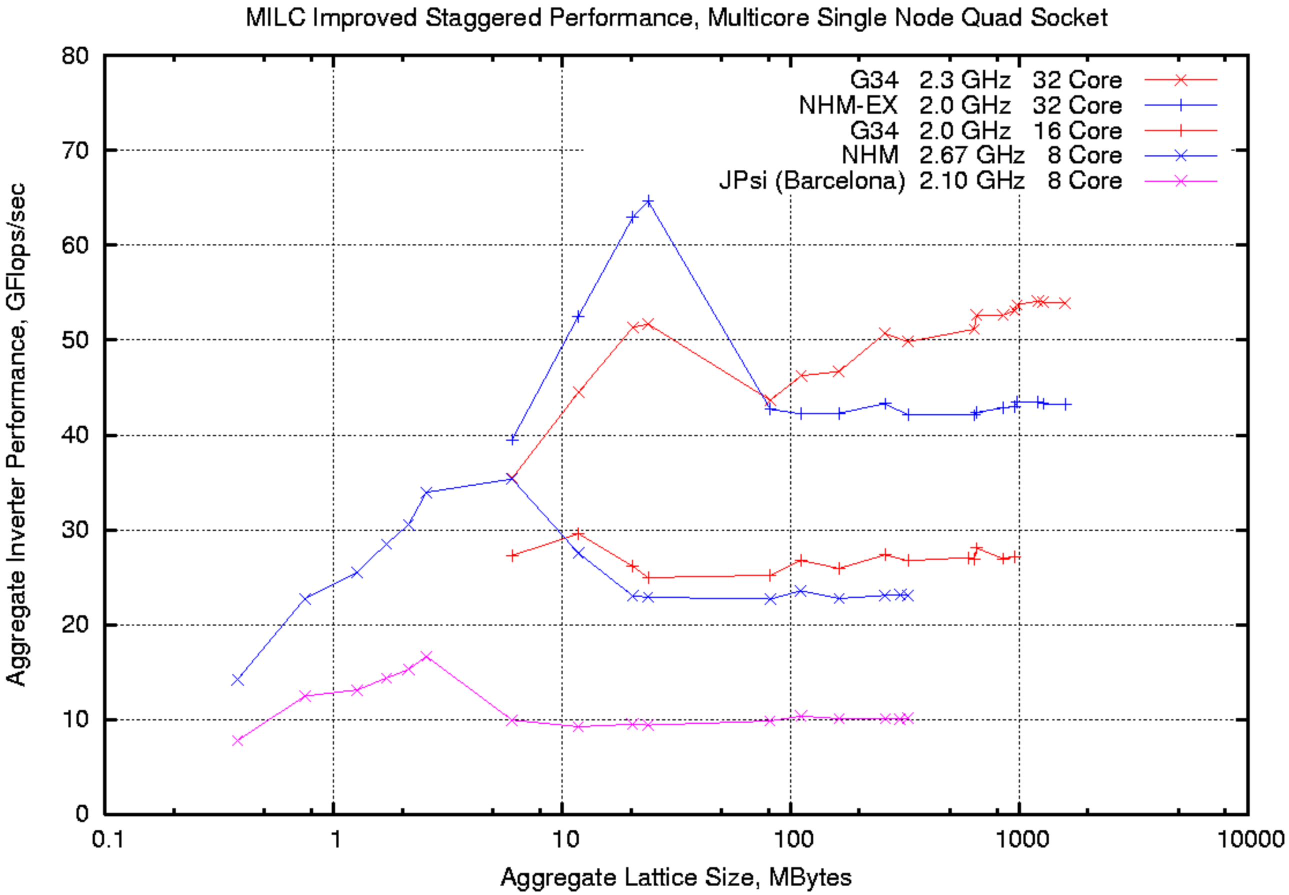
LQCD Cluster Layout



“Ds” Details

- Vendor: Koi Computers (Lombard, Illinois)
- Cost: $1.43M
- Nodes:
  - Quad socket, 8 cores/socket, 2.0 GHz, AMD “Magny-Cours” processors
  - 64 GBytes memory per node, 250 GB local disk
  - 21 nodes per rack, 15 KW maximum per rack (208V, 72A)
- Networking:
  - Quad data rate Infiniband (Mellanox)
  - 40 Gbits/sec/direction signaling rate (32 Gbits/sec data rate)
- Storage:
  - 258 TByte Lustre filesystem (expanding to 392 TBytes)
  - Shared with Kaon and J/Psi clusters

Performance of Candidate Processors



AMD Magny-Cours, 8 cores per socket. Top curve is 4 socket system, bottom curve is 2 socket system

Intel Nehalem EP and Nehalem EX. Top curve is 4 socket system, bottom curve is 2 socket system

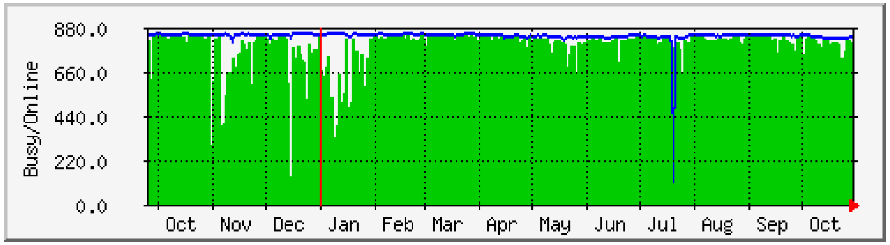
Reference: SC LQCD J/Psi cluster, AMD "Barcelona" 4 cores per socket

Four socket versions of Intel and AMD processors show essentially perfect scaling over two socket versions

Operations

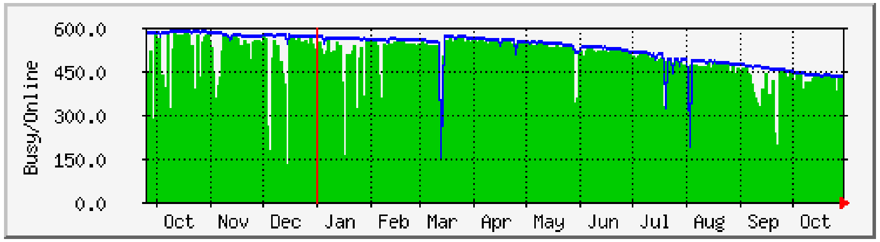
- LQCD-ext performance goals include:
  Delivered TFlops-yrs (uptime)
  Deployment TFlops (performance/price)
  Utilization (number of users, degree of use)
- Fermilab results (typical of all 3 labs):
  FY10 uptime = 98.8%
  FY10 deployment = 12.5 TF (goal = 11.0 TF)
  Utilization: 56 users, 90%+ utilization

`Yearly' Graph (1 Day Average)



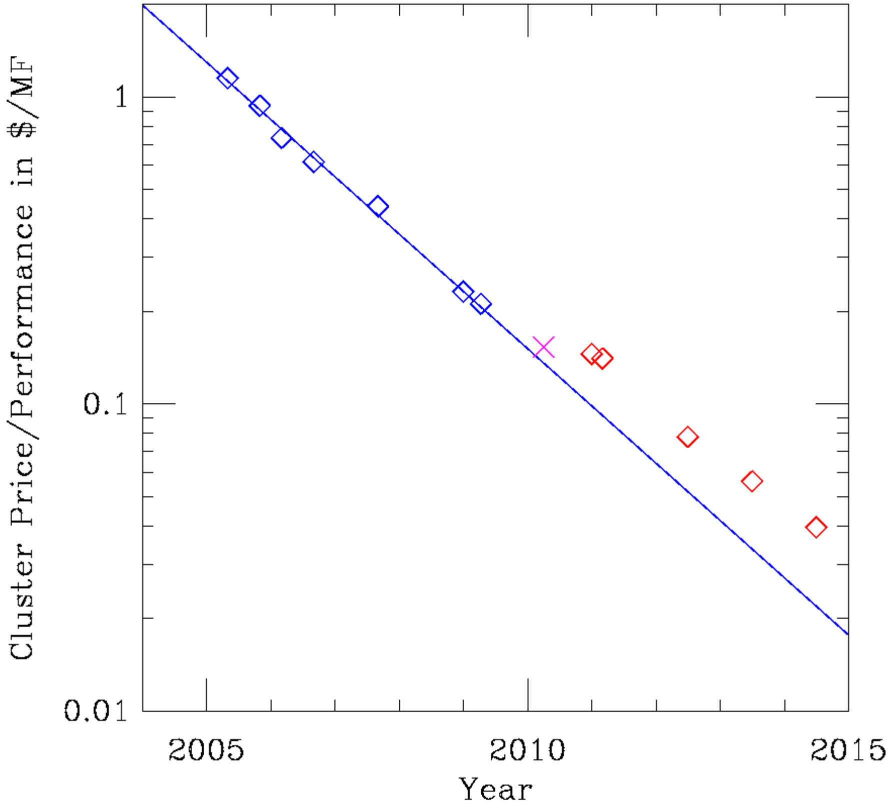
Max Busy: 851.0 Average Busy: 773.0 Current Busy: 776.0
Max Online: 856.0 Average Online: 845.0 Current Online: 839.0

`Yearly' Graph (1 Day Average)



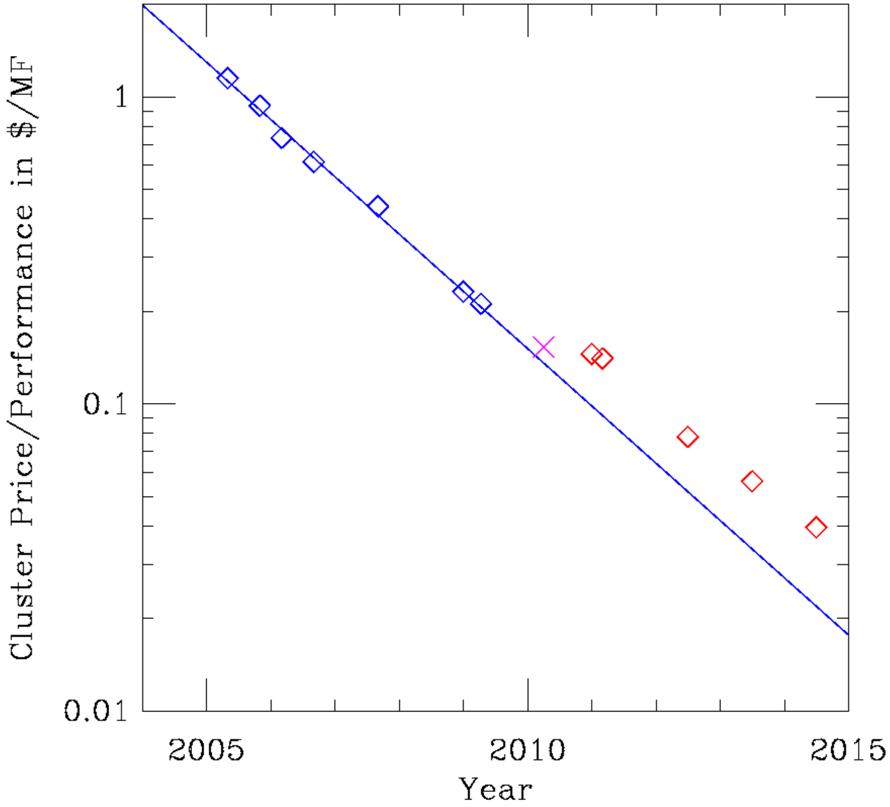
Max Busy: 586.0 Average Busy: 483.0 Current Busy: 408.0
Max Online: 591.0 Average Online: 533.0 Current Online: 434.0

Cost and Performance Basis



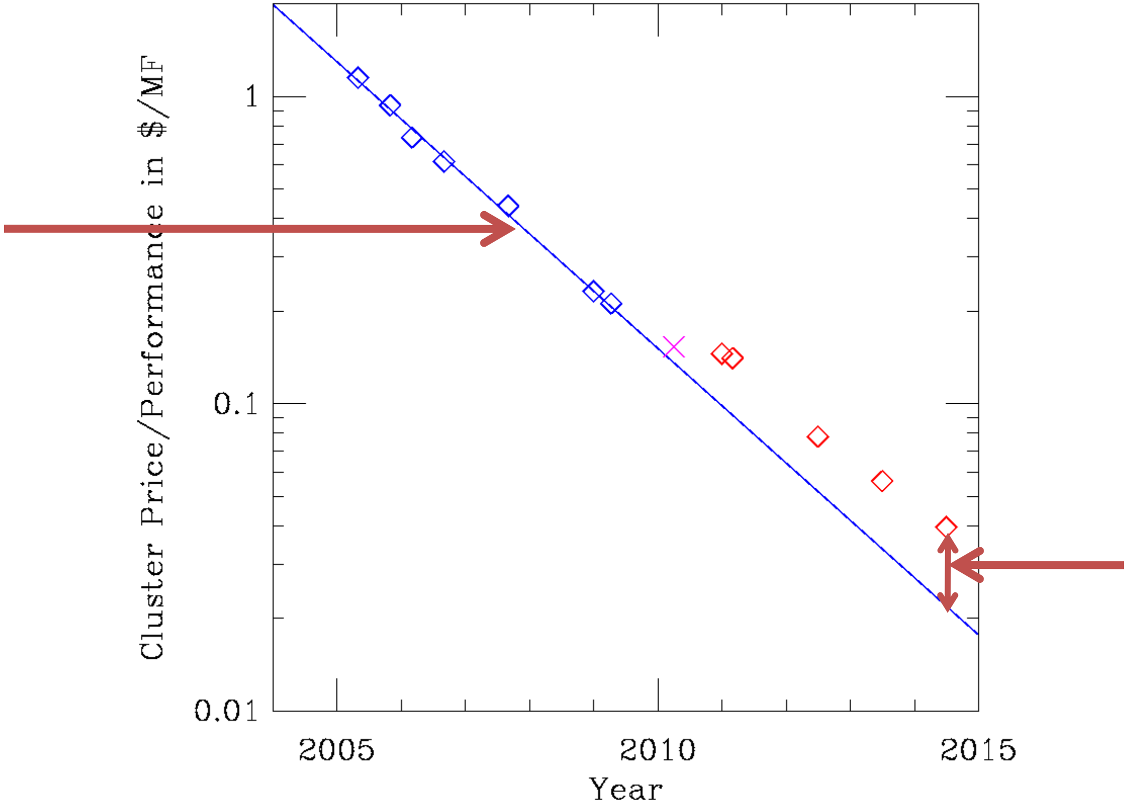
Cluster	Price per Node	Performance/Node, MF	Price/Performance
Pion #1	$1910	1660	$1.15/MF
Pion #2	$1554	1660	$0.94/MF
6n	$1785	2430	$0.74/MF
Kaon	$2617	4260	$0.61/MF
7n	$3320	7550	$0.44/MF
J/Psi #1	$2274	9810	$0.23/MF
J/Psi #2	$2082	9810	$0.21/MF
10q	$3461	22667	$0.15/MF

Cost and Performance Basis



Cluster	Price per Node	Performance/Node, MF	Price/Performance
Pion #1	$1910	1660	$1.15/MF
Pion #2	$1554	1660	$0.94/MF
6n	$1785	2430	$0.74/MF
Kaon	$2617	4260	$0.61/MF
7n	$3320	7550	$0.44/MF
J/Psi #1	$2274	9810	$0.23/MF
J/Psi #2	$2082	9810	$0.21/MF
10q	$3461	22667	$0.15/MF
Ds	$5810	50810	$0.114/MF

Cost and Performance Basis

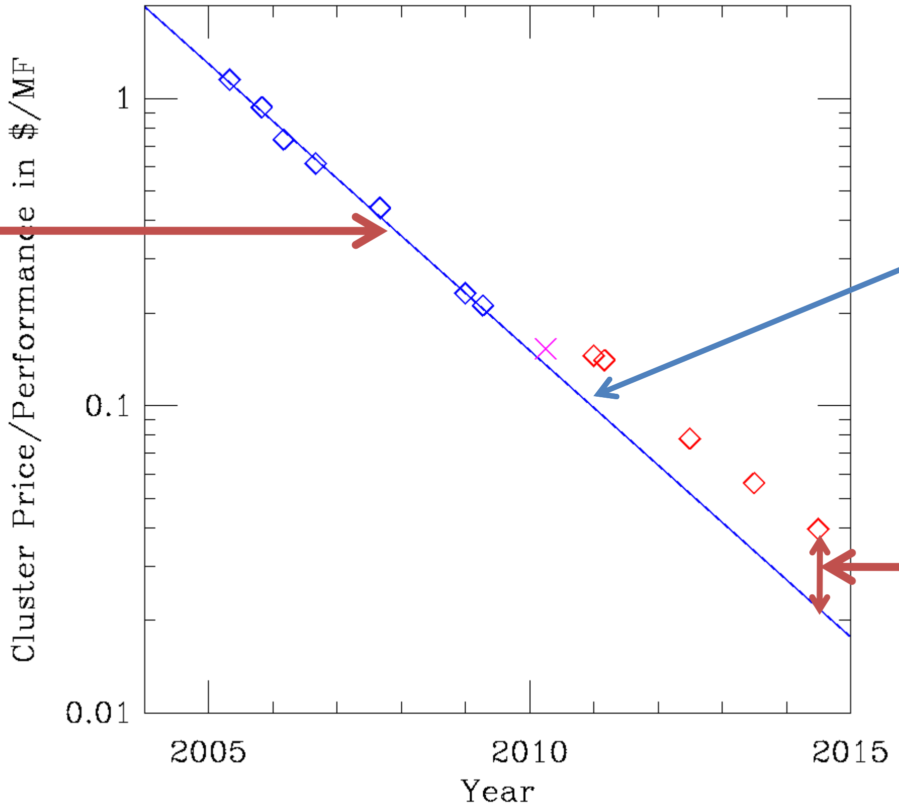


Fit is to the blue diamonds, slope gives halving time of 1.613 years

Contingency

Year	Deploy Date	Price/Perf. Goal	Price/Perf. Trend	Goal (TF)	Contingency (TF)	Contingency (TF %)
2010	2011.0	\$0.15/MF	\$0.098/MF	11	4.4	40%
2011	2011.2	\$0.14/MF	\$0.098/MF	12	4.4	36%
2012	2012.5	\$0.078/MF	\$0.052/MF	24	11.9	50%
2013	2013.5	\$0.056/MF	\$0.034/MF	44	26.8	61%
2014	2014.5	\$0.040/MF	\$0.022/MF	57	42.6	75%

Cost and Performance Basis



Fit is to the red diamonds, slope gives halving time of 1.613 years

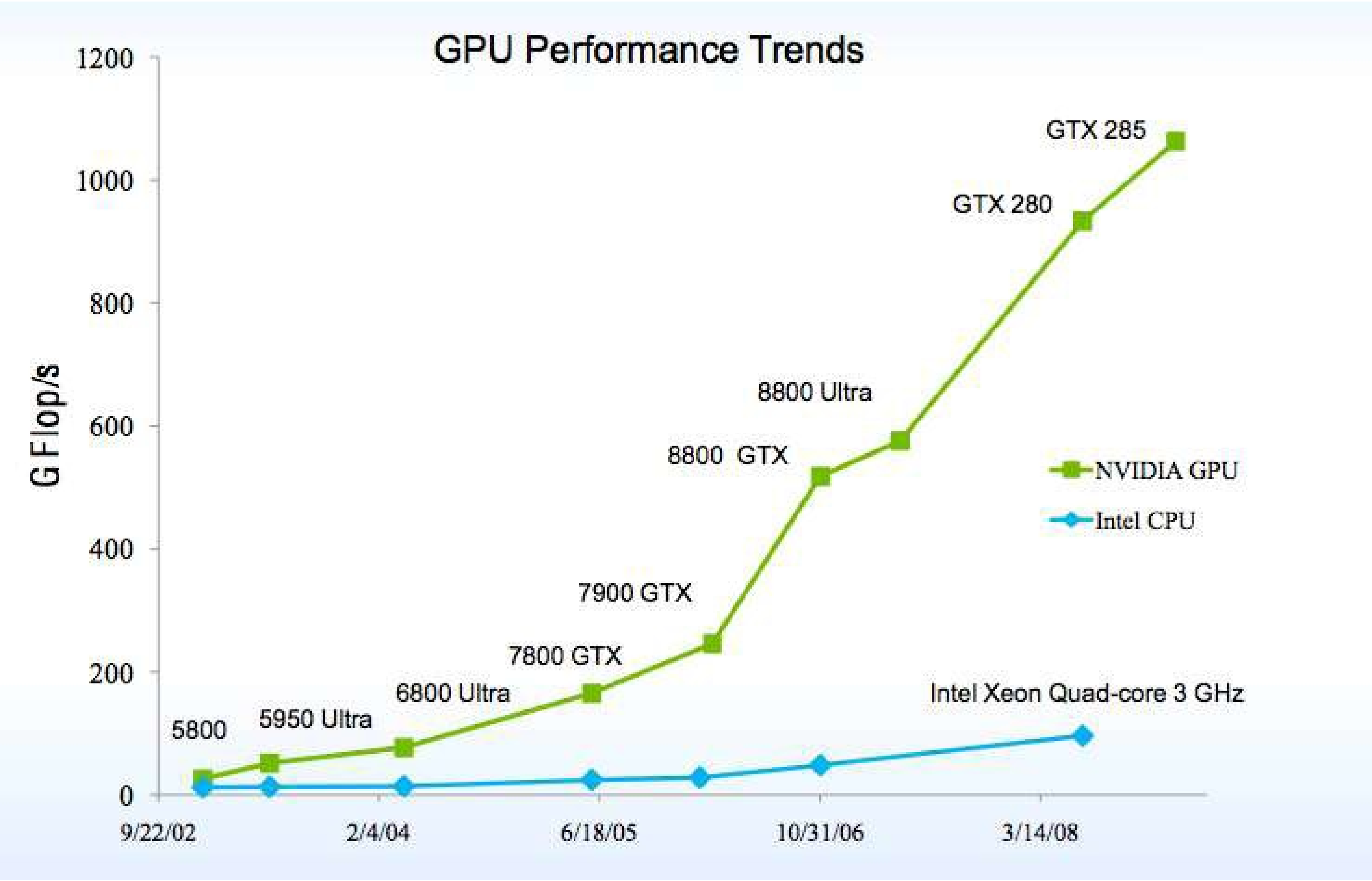
Ds: $0.114/MF

Contingency

Year	Deploy Date	Price/Perf. Goal	Price/Perf. Trend	Goal (TF)	Contingency (TF)	Contingency (TF %)
2010	2011.0	$0.15/MF	$0.098/MF	11	4.4	40%
2011	2011.2	$0.14/MF	$0.098/MF	12	4.4	36%
2012	2012.5	$0.078/MF	$0.052/MF	24	11.9	50%
2013	2013.5	$0.056/MF	$0.034/MF	44	26.8	61%
2014	2014.5	$0.040/MF	$0.022/MF	57	42.6	75%

FY11 Plans

- FY11 LQCD-ext budget for hardware = $1.69M
  - Purchase contract with Koi allows “Ds” expansion by up to 16 racks if ordered by March 31 (FY10 piece has 12 racks)
  - Project goal for FY11 is to deploy 12 TF
- FY11 budget will be split in some fraction between “Ds” expansion and a GPU cluster
  - GPUs are increasingly being exploited by lattice theorists
  - Up to a 10x performance gain on Dirac inverter (comparing single GPU to single J/Psi node)
  - GPU software development is very labor intensive
  - JLab ARRA LQCD FY09-10 cluster now in production with ~ 500 GPUs
  - Fermilab has 16 GPUs in production for LQCD, and 8 GPUs available for experimentation by any interested party



GPU Performance Trends